マルチモーダル情報を利用した英会話能力に関する研究

修士学位論文要約（平成31年３月）

マルチモーダル情報を利用した英会話能力に関する研究

呉　浩然

指導教員：伊藤 彰則，　研究指導教員：能勢　隆　千葉　祐弥

A Study on Evaluating English Proficiency Using Multimodal Information

Haoran WU

Supervisor: Akinori ITO,　Research Advisor: Takashi NOSE,　Yuya CHIBA

English proficiency is important for communication. Various systems are developed to help learners improve their English proficiency such as Computer-Assisted Language Learning (CALL) systems. Most of the conventional CALL systems concentrate on evaluating verbal fluency of the learners. However, not only verbal expressions but also facial expressions and gestures, are involved in actual English communication. However, it remains unknown how physical expressions affect the overall proficiency of English. Therefore, our study investigates the relationship between the fluency of the physical expression and the English proficiency by analyzing the multimodal dialog data, then an automatic evaluation of the learner's "facial expression / gesture" score is performed by multiple regression analysis. Results show high correlation with average evaluating score, which suggests the feasibility of using multimodal information in the assessment of English proficiency.

1. はじめに

近年のグローバル化に伴い，「発音」「聞き取り」のようなスキルを向上させるための言語学習支援[1](CALL, Computer Assisted Language Learning) システムが導入されるようになった．しかし実際の英会話においては，英語の発音だけでなく，会話に伴う表情やジェスチャなどを含めた総合的なコミュニケーション能力が求められている．しかしながら，表情やジェスチャなどの非言語的情報が英会話の流暢さの評価にどの程度影響するかは明らかではない．本研究では，マルチモーダル対話データを収集し，データに対して言語・非言語情報の尺度で主観評定実験を行い，表情，ジェスチャの自然性と会話の流暢さ及び自然性との関連性を分析する．また，機械学習による表情・ジェスチャの自然性の自動推定を行う．

2. マルチモーダル英会話データベース

本研究では，学習者が対話システムとシナリオベースで英会話を行う対話型 CALL システムを想定する．表情・ジェスチャなどが収録されるマルチモーダル英会話データベースを構築した．英文のシナリオとして，日本人向け英語教材から 2 つの対話シナリオを選定した．学習者はモニタに対面して座り，シナリオに沿ってシステムと対話を行った．このとき，ビデオカメラを学習者の正面に設置し，発話と動作を収録した．システムは実験者が操作し，学習者の回答が終了した時点で，次のシステム発話の動画を再生した．実験参加者は日本語母語話者の大学生と大学院生 13 名(男性 9 名，女性 4 名) であった．

3. 対話データの流暢さと自然性の評価

評定者には，日本人学生に対して教育経験を持つ，アメリカ国籍の英語語学教師男性 3 名を採用した．評定基準は五段階主観評定とした．本研究では表情・ジェスチャ，音素，韻律，総合スコアの 4 つの尺度に対して評定を付与した．各尺度の総合スコアへの寄与を標準化偏回帰係数で分析したところ，表情・ジェスチャの自然性が，音素に次いで 2 番目に大きいことが示された．これらの結果より，表情・ジェスチャは総合スコアに大きく影響することが示唆される．

4. 表情・姿勢特徴量の抽出

前節の分析結果は，表情・ジェスチャの自然性と総合スコアの関係性を示しているだけであり，具体的にどの情報が表情とジェスチャの評価に関わっているのかは明らかではなかった．従って画像データから表情・姿勢特徴量を抽出し，表情・ジェスチャの自然性の推定に有用な特徴量を分析する．

マルチモーダル特徴量の抽出には OpenFace [2] と OpenPose [3] を利用した．OpenFace は動画像中の顔の特徴点を含む視線情報や顔向き，Facial Action Units (AU) などの情報が網羅的に取得できる．一方で，OpenPose は各関節の座標によって表される姿勢情報が抽出できる．分析では，OpenFace と OpenPose から出力された各特徴量に対して系列ごとに計算された平均(avg) と分散(var) を利用した．これらの特徴量の統計量それぞれと「表情・ジェスチャ」に関する 3 名の評定値の平均の相関係数を求め，得られた特徴量のうち，「表情・ジェスチャ」の主観評定との相関係数

の絶対値が大きい特徴量上位 10 個を表 1 にまとめる.

表情特徴量(OpenFace)	相関係数	姿勢特徴量(OpenPose)	相関係数
視線方向(y 方向)_avg	0.507	身体(8)_x_var	0.372
頭部方向(pitch)_avg	0.482	左手(19)y_avg	0.361
AU12(唇両端を引上げる)_avg	0.464	左手(0)_x_var	0.353
AU45(まばたく)_r_avg	0.414	左手(2)_x_var	0.337
AU12(唇両端を引上げる)_var	0.381	左手(20)_x_var	0.323
AU07(瞼を緊張させる)_avg	0.361	左手(3)_x_var	0.323
AU17(オトガイをあげる)_avg	0.358	左手(1)_x_var	0.322
AU14(えくぼを作る)_avg	0.356	左手(19)_x_var	0.315
AU17(オトガイをあげる)_var	0.339	右手(12)_x_var	0.302
頭部位置(z 方向)_avg	0.331	左手(17)_x_var	0.299

表 1　OpenFace と OpenPose で抽出した特徴量の統計量のうち評定スコアとの相関が大きいもの(それぞれ上位 10 個)

5.表情・姿勢特徴量の分析

表 1 より, 特に視線や頭部方向の上下の動きの平均値が主観評定と高い相関を持つことがわかる. これは, 評定の低い学習者が頻繁に上方向を見ていることに起因している. このような学習者は, 会話をしながら発話内容を思い出しているように見えることから, 自然性を低く評価される傾向にあると考えられる. また, 興味深い結果として唇両端が引き上げられた時に値が大きくなる AU12 と主観評価との相関も比較的高い値を示している. これは, 評定の高い学習者が笑顔に近い表情で会話していることを反映している.

一方で, OpenPose で得られた姿勢特徴量の統計量は表情特徴量に比べると全体的に相関が小さい傾向にあることがわかる.

6.表情・ジェスチャスコアの自動評定

前節の結果により,「表情・ジェスチャ」などのマルチモーダル情報は英会話学習者の発話流暢さに大きく影響していることが分かった. 従って, 従来の発音と文法だけでなく, 学習者の「表情・ジェスチャ」などのマルチモーダル情報への自動評定, フィードバック可能な CALL システムが望まれる. 従って, 従来の発音と文法だけでなく, 学習者の「表情・ジェスチャ」などのマルチモーダル情報への自動評定, フィードバック可能な CALL システムが望まれる.

表情・ジェスチャスコアの自動評定に当たって, 利用する特徴量を選出し, 適切な学習手法を用いて, 機械学習させる必要がある. 本論文では, OpenPose と OpenFace で抽出した特徴量のうち, それぞれ「表情・ジェスチャ」の主観評定との相関係数の絶対値上位 10 個の特徴量を利用し, 重回帰分析で自動推定を行った. 本実験における重回帰モデルは:

1. OpenFace で抽出した特徴量(「表情・ジェスチャ」主観評定スコアとの相関係数の絶対値上位 10 個)
2. OpenPose で抽出した特徴量(「表情・ジェスチャ」主観評定スコアとの相関係数の絶対値上位 10 個)
3. OpenFace と OpenPose で抽出した特徴量(「表情・ジェスチャ」主観評定スコアとの相関係数の絶対値上位 16 個)

それぞれの条件で回帰分析モデルの決定係数, P 値などを算出し, leave-One-Out で予測されたスコアと実際のスコアとの相関係数を計算した. 表 2 にまとめる. OpenFace と OpenPose で抽出した特徴量を両方利用し Leave-One-Out で得た予測スコアシリーズと実際スコアシリーズとの相関は 0.458 であり, 比較的に高い相関を持っていることが示唆された.

実験条件	相関係数
OpenFace	0.364
OpenPose	0.125
OpenFace ,OpenPose	0.458

表 2: 予測スコアと実際スコアの相関係数

7. まとめ

本研究では, 言語学習支援 (CALL) システムの発展, 現状と欠点を述べ, マルチモーダル英会話学習支援 (CALL) システムの必要性を述べた. マルチモーダル英会話 CALL システムへの応用を想定し, マルチモーダルデータベースを収録した. 特に, 先行研究において考察が不十分であった「表情・ジェスチャ」の収録を行った. 次に, 収録したマルチモーダルデータベースを評定者に各尺度にて評定してもらい, 評定結果の有効性と各尺度の重要性の分析を行った. また, 機械で抽出したマルチモーダル情報と主観評定結果の関連性を検討した. また, 機械学習により「表情・ジェスチャ」の自動評定実験を行い, OpenFace と OpenPose で抽出した特徴量を両方利用し Leave-One-Out で得た予測スコアシリーズと実際のスコアシリーズとの相関は最も高いことが分かった.

文献

1) 河原他, "音声情報処理技術を用いた外国語学習支援, "電子情報通信学会論文誌, J96-D(7), 549-1565, 2013.
2) Baltrusaitis et al., "OpenFace: an open source facial behavior analysis toolkit." In Proc. IEEE Winter Conference on Applications of Computer Vision, 1-10, 2016.
3) Cao et al., "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," arXiv preprint arXiv: 1611.08050, 1-9, 2016.